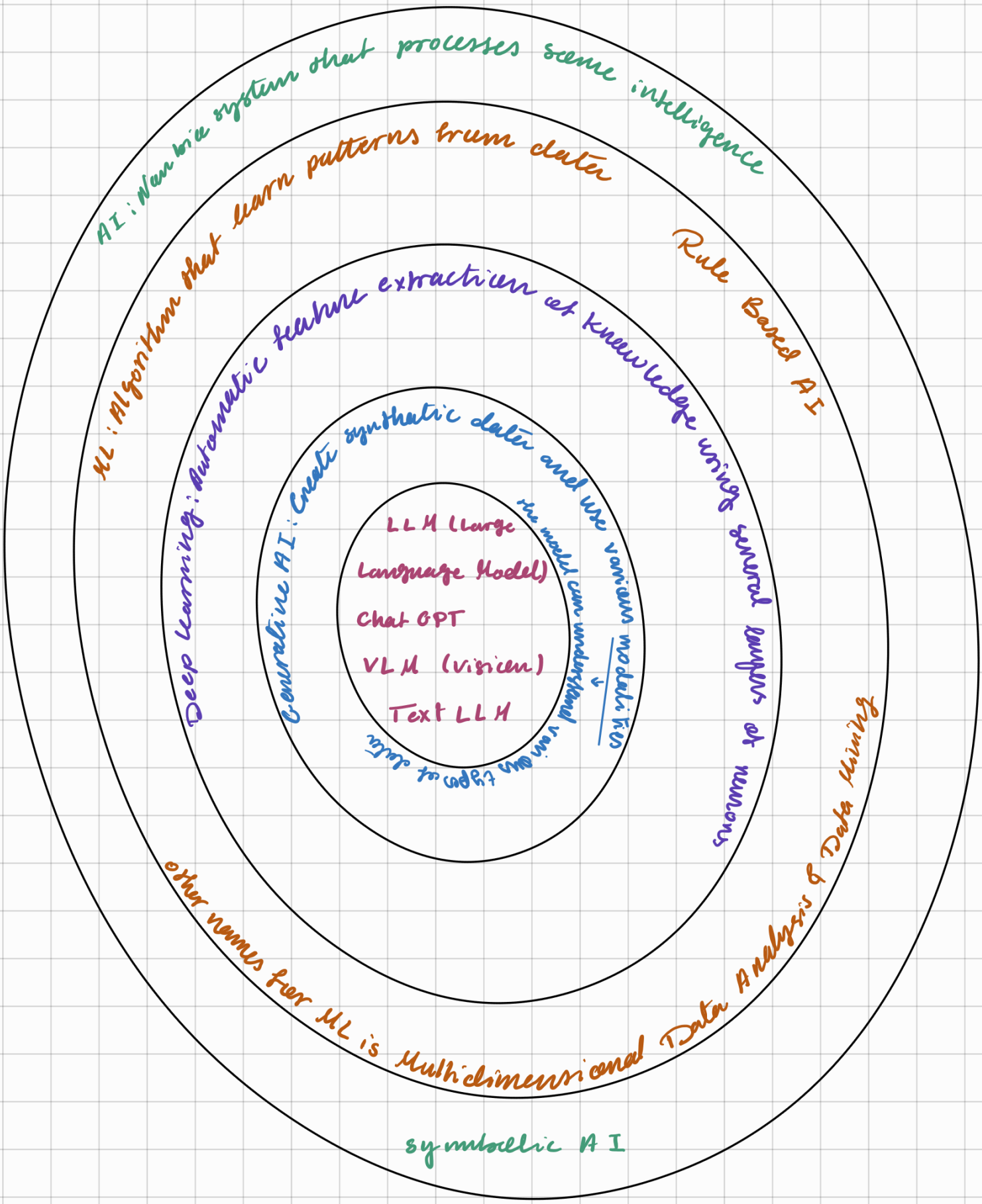
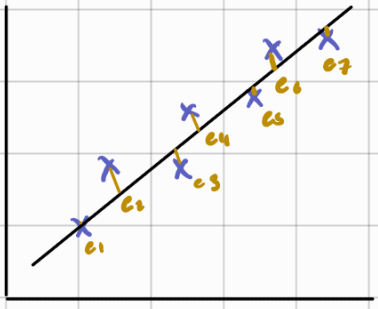


Note : Cornell University (put papers on public domain)
Learn Prompt engineering Coursera

Chat GPT (Generative Pre-trained Transformer Model)

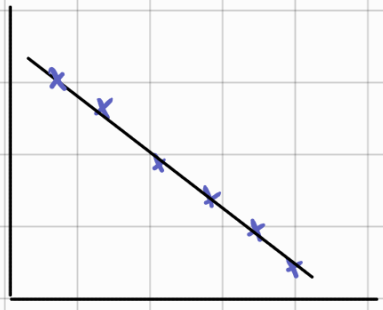
Chat GPT (Generative Pre-trained Transformer model) is a LLM that generates text.

Chat GPT is an optimization model that modifies the answers slightly.



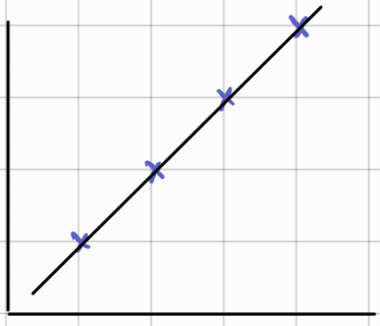
equation of the model $\hat{y} = \hat{\beta} x + \hat{\alpha}$

$\min\left(\sum_{i=1}^{n} e_i^2\right)$

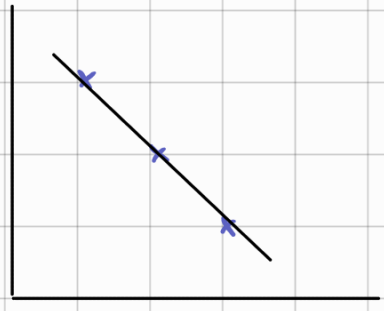


$\rho_{xy} \approx 0.8 > 0$

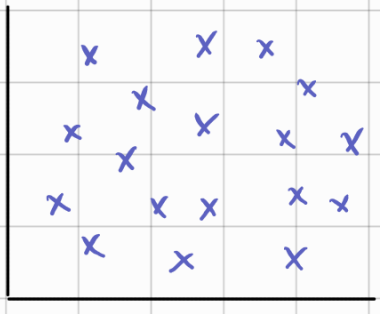
$\rho_{xy} \approx -0.8 < 0$



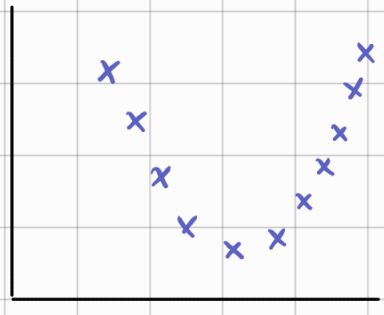
$\rho_{xy} = 1 > 0$



$\rho_{xy} = -1 < 0$



$\rho_{xy} = 0$



$\rho_{xy} = 0$ (non linear)

$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var\,x \cdot Var\,y}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Linear Regression assumes that the relation between the

variables is linear $y = \beta x + \alpha + \varepsilon \rightarrow$ error

Real Life Events Comprise of 1) Uncertainty
2) Variability

$y_i = \alpha + \beta x_i + \varepsilon_i$, error distribution always follows normal iid

α: intercept, β: slope, x_i: independent variable, ε_i: error

Instead of minimizing ε_i, we minimize $\sum e_i^2$

Sum of Squared Error: $SS = \min \left(\sum_{i=1}^{n} (y_i - \beta x_i - \alpha)^2 \right)$

$$\frac{\partial SS}{\partial \alpha} \overset{set}{=} 0 \qquad \frac{\partial SS}{\partial \beta} \overset{set}{=} 0$$

See we get optimal: $\hat{\alpha}$ and $\hat{\beta}$

$y_i = \hat{\beta} x_i + \hat{\alpha}$

For now, we assume: $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{Cov(x,y)}{Var(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$

x_i	$\bar{x}$	y_i	$\bar{y}$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3	1	2	-2	-1	2
2		1		-1	-1	1
3		2		0	0	0
4		2		1	0	0
5		4		2	2	4

$\hat{\beta} = \frac{2+1+4}{4+1+1+4} = \frac{7}{10}$ $\qquad \hat{\alpha} = 2 - \frac{7}{10} \cdot 3 = \frac{-1}{10}$

$y_i = 0.7 x_i - 0.1$ is the best fitted line!

* Regression is a tool for investigating the relationship between a dependent variable and one or more indep. variables. It has applications in economics, AI, health, engineering, etc. Before investigating, we study the correlation and plot the scatter plot to study the extent of the linear association.

Simple regression is used to study and model the relationship between one single regressor X and one response Y.

X : Regressor or control, not RV, has uncertainty & variability (regressing y on x).

Y : Dependent random variable

Equation of simple linear regression: $y = \beta_1 x + \beta_0 + \varepsilon$
where ε is the random error, β_0 intercept & β_1 slope.

Assumption: Both X and Y are continuous i.e they cannot be $\{0, 1\}$ but can be $[0, 1]$

* Physical interpretation of β_0 and β_1: $\hat{Y} = \beta_0 + \beta_1 X$

we think of $\hat{Y}$ as the profit or price and its unit is y, x's unit is x so the unit of β_0 is y and the unit of β_1 is y/x. We usually think of X as the investment.

* Given the data $(x_i, y_i)_{1 \le i \le n}$, we check the scatter plot to find whether the linear model is appropriate or not (if not change the model).

Simple linear regression model assumptions :

1) Homoscedasticity Assumption :
ε_i is a random variable with mean 0 and sd = σ unknown
i.e $\varepsilon_i \sim N(0, \sigma^2)$, $E(\varepsilon_i) = 0$ $\quad Var(\varepsilon_i) = \sigma^2$

2) $Cov(\varepsilon_i, \varepsilon_j) = 0 \Rightarrow \varepsilon_i$ and ε_j are independent

3) $\varepsilon_i \overset{\text{indep}}{\sim} N(0, \sigma^2)$

stochastic variable ↑ $\quad$ non-stocastic variable ↑

We declare that : $E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i$

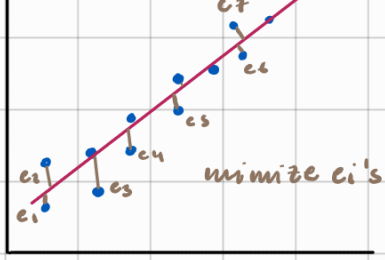
$Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$

$Y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$

Then, if Y_i is not normal, the model fails.

* Best Fitted β_0 and β_1 :

The line fitted by the least squares is the one that makes the sum of squares of the vertical discrepancies as small as possible. Our aim is to estimate β_0 and β_1 so that the sum of squares of all the differences between the observations Y_i and the line (fitted) is minimum.



Sum of squared errors (SS) :

$$S = \min\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) = \min\left(\sum_{i=1}^{n} e_i^2\right)$$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i is estimated by e_i (residual error)

y_i	$\hat{y}_{iD}$	$\hat{y}_{iz}$
-	-	-
-	-	-

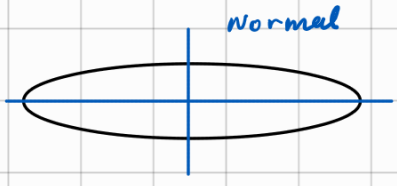
The best is $\min\{\sum(y_i - \hat{y}_{iD})^2, \sum(y_i - \hat{y}_{iz})^2\}$

Residual Error : $e_i = y_i - \hat{y}_i$

(y_i : true response, $\hat{y}_i$: estimated response)

* Least square estimator of β_0 and β_1 must satisfy :

Normal equations :



$$\begin{cases} \left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (1) \\ \left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \quad (2) \end{cases}$$

So $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions of the normal equations.

* Solving the equations by Ordinary least square (OLS) :

1) $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0$

$\Leftrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

2) $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i)x_i = 0$

$\Leftrightarrow \sum_{i=1}^{n}(y_i - \bar{y})x_i - \hat{\beta}_1(x_i - \bar{x})x_i = 0$

$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) =$

$\Leftrightarrow \sum_{i=1}^{n}(y_i - \bar{y})x_i = \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})x_i$

$\sum_{i=1}^{n}(y_i - \bar{y})x_i - \bar{x}\sum_{i=1}^{n}(y_i - \bar{y}) =$

$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i - n(n\bar{y} - n\bar{y})$$

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i \quad \Leftrightarrow \quad \hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\Leftrightarrow \quad \hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} = \frac{S_{xy}}{S_{xx}} \quad \begin{matrix} \rightarrow \text{without } \frac{1}{n} \\ \rightarrow \text{without } \frac{1}{n} \end{matrix}$$

$$\sum_{i=1}^{n} (y_i - \bar{y}) \bar{x} = \bar{x} \sum_{i=1}^{n} y_i - \bar{x}\bar{y}n = n\bar{x}\bar{y} - n\bar{x}\bar{y} = 0$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) \bar{x} = \bar{x} \sum_{i=1}^{n} x_i - n\bar{x}^2 = n\bar{x}^2 - n\bar{x}^2 = 0$$

So we added a zero term!

* Properties of Least Square Fit :

1) $\sum_{i=1}^{n} e_i = 0$ — sum of residuals in linear model = 0 comes from the 1st normal equal

2) $\sum_{i=1}^{n} e_i x_i = 0$ — sum of product of residual with regressor = 0 comes from the 2nd normal equal

3) $\sum_{i=1}^{n} e_i \hat{y}_i = 0$ — but $\sum_{i=1}^{n} e_i y_i \neq 0$

* What if we use MLE method ?

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon_i^2}{2\sigma^2}\right)$$

$$L(\varepsilon_i) = \prod_{i=1}^{n} f(\varepsilon_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon_i^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2}$$

$$L(\varepsilon_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$L^* = \log(L) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial L^*}{\partial \beta_0} = \frac{-2}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L^*}{\partial \beta_1} = \frac{-2}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

$$\frac{\partial L^*}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\Rightarrow \begin{cases} \hat{\beta}_{0\,MLE} = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_{1\,MLE} = S_{XY} / S_{YY} \\ \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{cases}$$

different than σ^2_{LSE}

$$\hat{\sigma}^2 = E\left(\frac{SS_{Res}}{n-2}\right)$$

* Both $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear estimators because :

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \sum_{i=1}^{n} c_i Y_i$, $c_i = \dfrac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

* Both estimators are unbiased as : measure of the goodness

of the estimators

$$E(\hat{\beta}_1) = E\left(\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\right)$$

$$= \frac{\sum (x_i-\bar{x})\, E(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

$$= \frac{\beta_1 \sum (x_i-\bar{x})^2}{\sum (x_i-\bar{x})^2}$$

$$= \beta_1$$

$$E\left(\frac{\sum (x_i-\bar{x})(\beta_0+\beta_1 x_i+\varepsilon_i-\bar{Y})}{\sum (x_i-\bar{x})^2}\right)$$

$$E\left(\frac{\sum (x_i-\bar{x})(\beta_0+\varepsilon_i-\bar{Y})+\beta_1\sum (x_i-\bar{x})x_i}{\sum (x_i-\bar{x})^2}\right)$$

$$\beta_1\, E\left(\frac{\sum (x_i-\bar{x})^2}{\sum (x_i-\bar{x})^2}\right)$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$$

$$(Y_i-\bar{Y}) = \beta_1 (x_i-\bar{x}) + (\varepsilon_i-\bar{\varepsilon})$$

$$E(Y_i-\bar{Y}) = \beta_1 (x_i-\bar{x}) + 0 \text{ as } \varepsilon_i \sim N(0,\sigma^2) \text{ so } \bar{\varepsilon} \sim N(0,\sigma^2/n)$$

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\beta_0 + \beta_1 \bar{x} - \hat{\beta}_1 \bar{x})$$

$$= \beta_0 + \beta_1 \bar{x} - \bar{x}\, E(\hat{\beta}_1)$$

$$= \beta_0 + \beta_1 \bar{x} - \bar{x}\,\beta_1$$

$$= \beta_0$$

$$* \quad Var(\hat{\beta}_1) = Var\left(\frac{\sum (x_i-\bar{x})(Y_i-\bar{Y})}{\sum (x_i-\bar{x})^2}\right)$$

$$= Var\left(\sum_{i=1}^{n} c_i Y_i\right)$$

$Y_i \overset{iid}{\sim} N(\beta_0+\beta_1 x_i, \sigma^2)$ ↙ so $Cov(Y_i, Y_j) = 0$

$$= \sum_{i=1}^{n} c_i^2\, Var(Y_i)$$

$$= \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} \cdot \sigma^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sigma^2}{S_{xx}}$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{X})$$

$$= \text{Var}(\bar{Y}) + \bar{X}^2 \text{Var}(\hat{\beta}_1) - 2\,\text{Cov}(\bar{Y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{X}^2 \cdot \frac{\sigma^2}{S_{xx}} - 2\,\text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} c_i Y_i\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right) - \frac{2}{n}\sum_{i=1}^{n} c_i \,\text{Var}(Y_i)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right) - \frac{2\sigma^2}{n}\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right) - \frac{2\sigma^2}{n}\left(\frac{n\bar{x} - n\bar{x}}{\sum (x_i - \bar{x})^2}\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)$$

* Estimation of σ^2 :

$E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2$, $\frac{SS_{Res}}{n-2}$ is an unbiased estimator of σ^2

Residual : $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2$$

$$= \sum_{i=1}^{n} \left((Y_i - \bar{Y})^2 + \hat{\beta}_1^2 (x_i - \bar{x})^2 - 2\hat{\beta}_1 (x_i - \bar{x})(Y_i - \bar{Y})\right)$$

$$= S_{YY} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xY}$$

$$= S_{YY} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1^2 S_{xx}$$

$$= S_{YY} - \hat{\beta}_1^2 S_{xx}$$

$$\sum(\beta_0+\beta_1 x_i)^2 - n(\beta_0+\beta_1\bar{x}) = n\beta_0^2 + 2\beta_0\beta_1 n\bar{x} + \beta_1^2\sum x_i^2 - n\beta_0^2 - 2n\beta_0\beta_1\bar{x} - n\beta_1^2\bar{x}^2$$

$$\sum(x_i-\bar{x})^2 = = \beta_1^2\left(\sum x_i^2 - \bar{x}^2\right)$$

$$\sum x_i^2 - \sum 2x_i\bar{x} + \sum\bar{x}^2 = = \beta_1^2\sum(x_i-\bar{x})^2$$

$$\sum x_i^2 - 2n\bar{x} + n\bar{x}^2 = = \beta_1^2 S_{xx}$$

$$\sum x_i^2 - \bar{x}^2$$

$$\begin{aligned}
E(SS_{RES}) &= E(S_{YY} - \hat{\beta}_1^2 S_{xx}) \\
&= E(S_{YY}) - S_{xx}E(\hat{\beta}_1^2) \\
&= E\left(\sum_{i=1}^n (Y_i-\bar{Y})^2\right) - S_{xx}\left(Var(\hat{\beta}_1) + E(\hat{\beta}_1)^2\right) \\
&= E\left(\sum_{i=1}^n (Y_i^2+\bar{Y}-2Y_i\bar{Y})\right) - S_{xx}\left(\sigma^2/S_{xx} + \beta_1^2\right) \\
&= E\left(\sum_{i=1}^n Y_i^2 + n\bar{Y} - 2\bar{Y}\sum_{i=1}^n Y_i\right) - \sigma^2 - \beta_1^2 S_{xx} \\
&= \sum_{i=1}^n E(Y_i^2) - nE(\bar{Y}^2) - \sigma^2 - \beta_1^2 S_{xx} \\
&= \sum_{i=1}^n (Var(Y_i) + E(Y_i)^2) - n(Var\,\bar{Y} + E(\bar{Y})^2) - \sigma^2 - \beta_1^2 S_{xx} \\
&= n\sigma^2 + \sum_{i=1}^n (\beta_0+\beta_1 x_i)^2 - \sigma^2 - n(\beta_0+\beta_1\bar{x})^2 - \sigma^2 - \beta_1^2 S_{xx} \\
&= \sigma^2(n-1) + \beta_1^2\left(\sum_{i=1}^n (x_i-\bar{x})^2\right) - \sigma^2 - \beta_1^2 S_{xx} \\
&= \sigma^2(n-1) + \beta_1^2 S_{xx} - \sigma^2 - \beta_1^2 S_{xx} \\
&= \sigma^2(n-2)
\end{aligned}$$

Hence, $E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2$, $MS_{Res} = \frac{SS_{Res}}{n-2}$

* What is the sampling distribution of MS_{Res}?

$Y_i \overset{iid}{\sim} N(\beta_0+\beta_1 x_i, \sigma^2)$ $\quad$ $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$$SS_{Res} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i-\hat{Y}) = \sum_{i=1}^n \left(\frac{Y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2 \sim \chi^2_{n-2}$$

so $\frac{e_i^2}{\sigma^2} \sim \chi^2_1$ and $\sum_{i=1}^n \frac{e_i^2}{\sigma^2} \sim \chi^2_{n-2}$ restrictions: $\sum e_i = 0$

normal equations $\sum e_i x_i = 0$

$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2} \Rightarrow \frac{SS_{Res}/n-2}{\sigma^2/n-2} = \frac{MSR}{\sigma^2}(n-2) \sim \chi^2_{n-2}$

$$MSR \sim \frac{\sigma^2}{n-2}\chi^2_{n-2}$$

* Statistic test of β_1 : Test of slope coefficient

we want to test if ∃ a linear relationship between x_i & y_i.

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

↓ ↓

no linear relationship ∃ a linear relationship

$y_i = \beta_0$

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \quad , \quad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$, we use z-test if σ is known

$$z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}$$, if $|z| > z_{\alpha/2}$ we reject H_0 (two sided)

If σ^2 is unknown : $t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}}$, reject if $|t| > t_{\alpha/2, n-2}$

↓ do not estimate σ^2

$$MS_{RES} = \frac{SS_{Res}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-2}$$

For β_0 : $$t^* = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{RES}}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

* **Example** : Sheew that $Cov(\hat{\beta}_0, \hat{\beta}_1) = 0$ when $\bar{x} = 0$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov\left(\bar{Y} - \hat{\beta}_1 \bar{x}, \sum_{i=1}^{n} c_i Y_i\right)$$

$$= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} Y_i\right)$$

$$= \frac{1}{n} \frac{\sum (x_i - \bar{x})}{S_{xx}} Cov(Y_i, Y_i)$$

$$= \frac{Var\, Y_i}{n\, S_{xx}} \left(\sum_{i=1}^{n} x_i - \bar{x}\right)$$

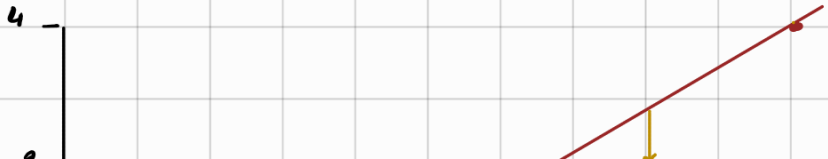
$$= \frac{\sigma^2}{n\, S_{xx}} (n\bar{x} - n\bar{x})$$

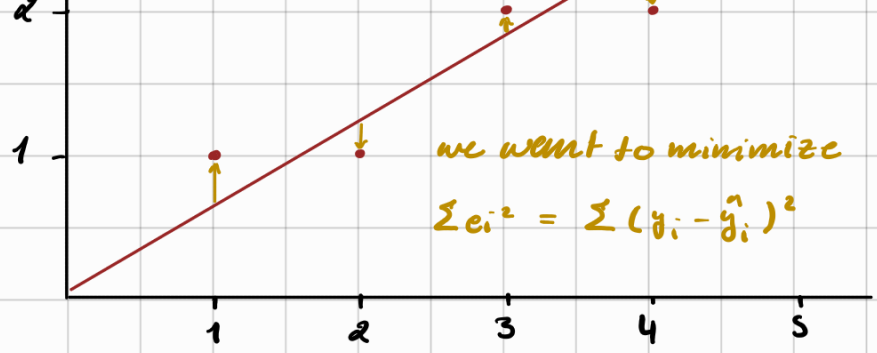
$$= 0$$

* **Example** : Suppose you have been hired by a chocolate company as a marketing analyst. They have the following data :

Advertisment cost (10K $)	Sales (10^5 units)
1	1
2	1
3	2
4	2
5	4

a) Is there a relationship between sales and advertisment ?





we want to minimize

$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$

There is a positive association between x_i and y_i as combined by the scatter plot and s_{xy}

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	S_{xx}	$Var\,x$	y_i	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	S_{yy}	$Var\,y$
1	3	-2	4	10	2	1	2	-1	1	6	$\frac{6}{5}$
2		-1	1			1		-1	1		
3		0	0			2		0	0		
4		1	1			2		0	0		
5		2	4			4		2	4		

$$s_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \frac{2+1+4}{\sqrt{10}\cdot\sqrt{6}} = \frac{7}{\sqrt{60}} = 0.9$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{7}{10} = 0.7 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - \frac{7}{10}\cdot 3 = -0.1$$

Model : $\hat{y} = -0.1 + 0.7x$

↓ ↓

non sense as companies 0.7 increase in sales per 1 \$ (10K \$) of investment sell units even without advertising, we need more data to estimate accurately.

b) Is the relationship significant at 5% level

x_i	y_i	$\hat{y}$	$e_i = y_i - \hat{y}$	e_i^2	$\sum e_i^2$

$$MS_{RES} = \frac{SS_{RES}}{}$$

1	1	0.6	0.4	0.16	1.1
2	1	1.3	-0.3	0.09	
3	2	2	0	0	
4	2	2.7	-0.7	0.49	
5	4	3.4	0.6	0.36	

$n-2$

$= \frac{1.1}{3}$

$= 0.364$

$H_0: \hat{\beta}_1 = 0$ vs $H_1: \hat{\beta}_1 \neq 0$ twice sided t-test

$t_c = t_{\alpha/2, n-2} = t_{0.025, 3} = 3.18$, we reject H_0 if $t > |t_c|$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSRes / S_{xx}}} = \frac{0.7}{\sqrt{\frac{0.364}{10}}} = 3.656$$

Hence, we reject H_0 and thus there is a significant relationship at 5% level.

$$\text{For } \hat{\beta}_0: \quad t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{RES}}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} = \frac{-0.1}{\sqrt{0.364}\sqrt{\frac{1}{5} + \frac{3^2}{10}}} = -0.1576$$

* To get the p-value : $\mathbb{P}(>|t|) = 2\,\mathbb{P}(> 3.18)$

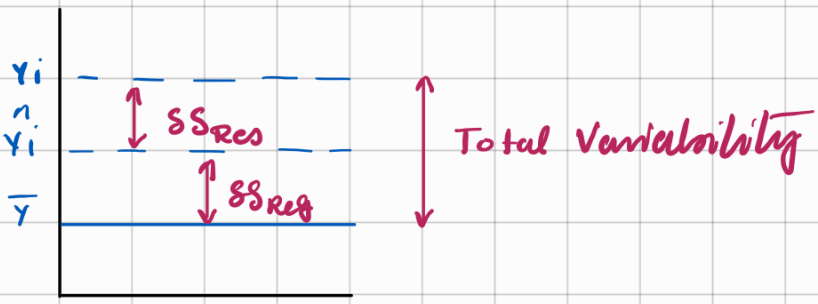
1) Look for $df = n-2$
2) Check the closest two values to the calculated t in row $n-2$
4) Take the average of their respective α
5) Multiply by 2.

$$\text{For } \hat{\beta}_1: \quad 2 \cdot \left(\frac{\overset{3.482}{0.02} + \overset{3.896}{0.015}}{2} \right) = 0.035 < 0.05, \text{ we reject } H_0$$

$$\text{For } \hat{\beta}_0: \quad 2 \cdot \left(\frac{0.45 + 0.40}{2} \right) = 0.85 > 0.05, \text{ cannot reject } H_0$$

$$\text{standard error} = \frac{\text{estimate}}{\text{P value}} = \frac{0.7}{0.035} = 20$$

c) Build an Anova table to test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$



Total variability of the data $= \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$\begin{aligned} SS_T &= \sum (y_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= SS_{Res} + SS_{Reg} \end{aligned}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

$$\begin{aligned} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^{n} \hat{\beta}_1 (x_i - \bar{x}) \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right) \\ &= \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx} \\ &= \hat{\beta}_1 S_{xy} - \hat{\beta}_1 S_{xy} \quad , \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \\ &= 0 \end{aligned}$$

SS_{Res} : Deviation of the predicted value from the actual value.

SS_{Reg} : Deviation of the predicted value to the mean

$$SS_{Reg} = \sum (\hat{Y}_i - \bar{Y})^2$$
$$= \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2$$
$$= \sum \hat{\beta}_1^2 (x_i - \bar{x})^2$$
$$= \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$
$$= \hat{\beta}_1^2 S_{xx}$$
$$= \hat{\beta}_1 S_{xy}$$

$SS_{Res} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2 \sim \chi^2_{n-2}$ → normal equations

SS_T : $\sum (y_i - \bar{y}) = 0$ is the only restriction : $n-1$ DF

So, SS_{Reg} has only one degree of freedom $\sim \chi^2_1$

$E(MS_{Res}) = \sigma^2$

$E(MS_{Reg}) = E(\hat{\beta}_1^2 S_{xx}) = S_{xx} (Var(\hat{\beta}_1) + E(\hat{\beta}_1)^2)$

$E(\hat{\beta}_1 S_{xy})$ = $\beta_1 S_{xy}$

$= S_{xx} \left(\frac{\sigma^2}{S_{xx}} + \beta_1^2 \right)$

$= \sigma^2 + S_{xx} \beta_1^2$

ANOVA TABLE :

Source of Variability	DF	SS	MS	F - Value
Regression	1	4.9	4.9	$F = \frac{MS_{Reg}}{MS_{Res}} \sim F_{1,n-2}$
Residual	3	1.1	0.366	$F = \frac{4.9}{0.366} = 13.38$
Total	4	6		

$F > F_{1,n-2,0.05} = 10.13$ so we reject H_0 and thus β_1 is statistically significant and the model is significant

$$SST = \sum_{i=1}^{5} (Y_i - \bar{Y})^2 = S_{YY} = 6$$

$$SS_{RES} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 0.16 + 0.09 + 0.49 + 0.36 = 1.1$$

$$\Rightarrow SS_{Reg} = SS_T - SS_{RES} = 4.9$$

Alternatively, $SS_{Reg} = \sum (Y_i - \hat{Y}_i)^2$

$$= (0.6-2)^2 + (1.3-2)^2 + (2-2)^2 + (2.7-2)^2 + (3.4-2)^2 = 4.9$$

$$MSReg = \frac{SSreg}{1} = \frac{4.9}{1} = 4.9$$

$$MSRes = \frac{SSres}{3} = \frac{1.1}{3} = 0.366$$

d) Check the accuracy of the regression model

Measure of "Goodness of fit":

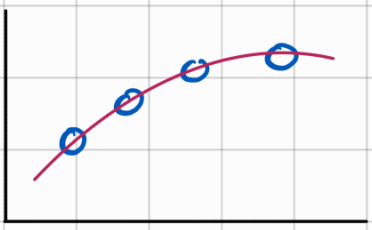
Coefficient of determination: R^2

R^2 measures the proportion of variability in response variable that is explained by the regression model.

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \quad \text{maximum } SS_{Reg} \text{ minimum } SS_{Res}$$

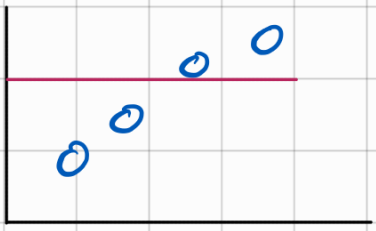
$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var\,x \; Var\,y}}$$

- $R^2 = 1$: $SS_T = SS_{Reg}$ Perfect fit



$SS_{Res} = 0$ so all the variability is explained by the model (var of the data)

- $R^2 = 0$: $SS_{Res} = SS_T \Rightarrow \sum (y_i - \hat{y})^2 = \sum (y_i - \bar{y})^2$
$\Rightarrow \hat{y}_i = \bar{y}$



No relationship so $\hat{\beta}_1 = 0$ then $S_{xy} = 0$

$R^2 = \frac{SS_{Res}}{SS_T} = \frac{\hat{\beta}_1^2 S_{xx} \neq 0}{SS_T} = 0 \Rightarrow \hat{\beta}_1 = 0$

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{4.9}{6} = 0.81\overline{6} = 0.82$$

Therefore, 82% of the total variability of the sales amount is explained by advertisment. It is not 100% because, it would an over fitted model and there might be some other variables that are useful to explain the remaining variability of the model but was not given e.g # of outlets & # of workers.

e) Find confidence interval for β_1

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, Normal as $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

S_{xx}

By CLT: $z = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSRes / S_{xx}}} \sim t_{\alpha, n-2} \quad , \quad \sigma^2 = \mathbb{E}(MSRes)$$

Thus, $\mathbb{P}\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{MSRes}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{MSRes}{S_{xx}}} \right) = 1 - \alpha$

$$\mathbb{P}\left(0.7 - t_{0.025,3} \sqrt{\frac{0.366}{10}} \leq \beta_1 \leq 0.7 + t_{0.025,3} \sqrt{\frac{0.366}{10}} \right) = 0.95$$

$\beta_1 \in \,]0.091, 1.31[$, $t_{0.025,3} = 3.182$

f) Estimate the mean sales amount when advertising cost is 4 (40K \$) at 0.05 level.

$\hat{Y} = -0.1 + 0.7 X$, $X = 4 \Rightarrow \hat{Y} = 2.7$

Interval estimation of mean response given $X = x_0$:

$\mathbb{E}(Y \mid X = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{Y}$

$$\begin{aligned} Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= Var(\bar{Y} + \hat{\beta}_1 (x_0 - \bar{x})) \\ &= Var\,\bar{Y} + (x_0 - \bar{x})^2 Var\,\hat{\beta}_1 - 2\,Cov(\bar{Y}, \hat{\beta}_0 (x_0 - \bar{x})) \\ &= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \cdot \frac{\sigma^2}{S_{xx}} - 2(x_0 - \bar{x})\, Cov(\bar{Y}, \hat{\beta}_0) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

($Cov(\bar{Y}, \hat{\beta}_0) = 0$: independent Y and β_0)

$$\mathbb{E}(Y \mid X = x_0) \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_0 , \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

$$\mathbb{E}(Y|X=x_0) \in \left]\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right.,$$

$$\left.\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right) + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right[$$

$$\mathbb{E}(Y|X=x_0) \in \left]2.7 - 3.812\sqrt{0.366\left(\frac{1}{5}+\frac{(4-3)^2}{10}\right)}, 2.7 + \right.$$

$$\left.3.812\sqrt{0.366\left(\frac{1}{5}+\frac{(4-3)^2}{10}\right)}\right[$$

$$\mathbb{E}(Y|X=x_0) \in \left]1.65, 3.75\right[$$

$$\mathbb{P}(1.65 \leq \mathbb{E}(Y|X=x_0) \leq 3.75) = 0.95$$